JOURNAL OF MULTIVARIATE ANALYSIS 55, 61–81 (1995)

Density Estimation under Qualitative Assumptions in Higher Dimensions

WOLFGANG POLONIK*

Universität Heidelberg, Heidelberg, Germany

We study a method for estimating a density f in $\mathbf{R}^d$ under assumptions which are of qualitative nature. The resulting density estimator can be considered as a generalization of the Grenander estimator for monotone densities. The assumptions on f are given in terms of shape restrictions of the density contour clusters $\Gamma(\lambda) = \{x : f(x) \geq \lambda\}$. We assume that for all $\lambda \geq 0$ the sets $\Gamma(\lambda)$ lie in a given class $\mathbb{C}$ of measurable subsets of $\mathbf{R}^d$. By choosing $\mathbb{C}$ appropriately it is possible to model for example monotonicity, symmetry, or multimodality. The main mathematical tool for proving consistency and rates of convergence of the density estimator is empirical process theory. It turns out that the rates depend on the richness of $\mathbb{C}$ measured by metric entropy.

1. INTRODUCTION

We study the problem of estimating a real valued function in $\mathbf{R}^d$ under qualitative assumptions like monotonicity, symmetry, or modality. In this paper we consider density estimation in an i.i.d. setup. Assumptions on the underlying density f are formulated as shape restrictions. We assume that all density contour clusters $\Gamma(\lambda) = \Gamma_f(\lambda) = \{x \in \mathbf{R}^d : f(x) \geq \lambda\}$, $\lambda > 0$, lie in a given class $\mathbb{C}$ of measurable subsets of $\mathbf{R}^d$. Under this assumption the sets $\Gamma(\lambda)$ are estimated by using the so-called excess mass approach. This leads to certain minimum volume sets (at random levels) as estimators for $\Gamma(\lambda)$. A minimum volume set in $\mathbb{C}$ at level α, by definition, has minimal Lebesgue measure among all sets in $\mathbb{C}$ containing empirical mass not less than α. Using these estimators of $\Gamma(\lambda)$ a density estimator called silhouette is constructed. The goal of this paper is to study the (asymptotic) behaviour of the silhouette.

Received September 16, 1993; revised March 1995

AMS 1990 subject classifications: primary 62G, secondary 62H.

Key words and phrases: *Grenander estimator, excess mass, density contour cluster, empirical process theory.*

* Research supported by the Sonderforschungsbereich 123 and the Deutsche Forschungsgemeinschaft.

0047-259X/95 $12.00

Maximum likelihood density estimation under shape restrictions has been considered in Robertson [16], Wegman [21], and Sager [19]. (See also the book of Robertson, Wright, and Dykstra [17]). They considered estimation of densities measurable with respect to σ-lattices. A σ-lattice (in $\mathbf{R}^d$) is a class of measurable subsets which is closed under countable unions and intersections and contains $\varnothing$ and $\mathbf{R}^d$. Measurability of f with respect to a σ-lattice $\mathbb{C}$ means $\Gamma(\lambda) \in \mathbb{C}\ \forall \lambda \geqslant 0$. The m.l.e. corresponding to the σ-lattice $\mathbb{M}_0 = \{[0, x],\ x \geqslant 0\}$ is the Grenander estimator of a monotone density on the positive real line (Grenander [6]). It is shown in Section 2 that the silhouette with $\mathbb{C} = \mathbb{M}_0$ equals the Grenander estimator. However, in the present paper $\mathbb{C}$ is not required to be a σ-lattice.

Other density estimators related to the silhouette have been considered by Sager [18, 20]. In [18] the sets $\Gamma(\lambda)$ are estimated by minimal volume sets in $\mathscr{C}^d$, where $\mathscr{C}^d$ denotes the class of closed convex sets in $\mathbf{R}^d$. Because minimum volume sets can be overlapping Sager constructed a *nested* sequence of minimal volume sets and used them to form a density estimator. In [20] Sager first estimates a real-valued function $x \to g(x)$, $x \in \mathbf{R}^d$, which is called isopleth form. (Typically $g(x)$ is the Lebesgue measure of the smallest density contour cluster containing x.) Then Sager uses the estimated isopleth form evaluated at the original d-dimensional data points as a new (one-dimensional) data set and estimates their (monotone) density which is called transfer density. The composition of these function estimators gives the final multivariate density estimator. For special $\mathbb{C}$ the silhouette is of the same structure (cf. Section 2).

Let us briefly discuss the model assumption $\Gamma(\lambda) \in \mathbb{C}\ \forall \lambda > 0$. It is equivalent to $f \in \mathscr{F}_{\mathbb{C}}$, where

$$\mathscr{F}_{\mathbb{C}} = \left\{ f\colon \mathbf{R}^d \to [0, \infty),\ \int f(x)\, dx = 1,\ \Gamma_f(\lambda) \in \mathbb{C}\ \forall \lambda > 0 \right\}.$$

With this notation $\mathscr{F}_{\mathbb{M}_0}$ is the class of all nonincreasing leftcontinuous densities on the positive real line. Let $\mathscr{I}_1$ denote the class of closed intervals. Then $\mathscr{F}_{\mathscr{I}_1}$ is the class of unimodal densities on the real line. In higher dimensions ($d \geqslant 2$) there is no such natural choice as the class of intervals for $d = 1$. Typical choices are the class $\mathscr{C}^d$ and the classes of all closed balls and closed ellipsoids in $\mathbf{R}^d$, denoted by $\mathscr{B}^d$, and $\mathscr{E}^d$, respectively. In the multimodal case density contours of a univariate density are unions of intervals. For $d \geqslant 2$ classes which can be constructed out of convex sets by means of finitely many set-theoretic operations $\cap$, $\cup$, ^{c} seem to be appropriate to model multimodality (see Polonik [13] for a discussion). Symmetry around zero corresponds to the class of all balls with midpoint zero. Note that the latter class and the class $\mathbb{M}_0$ are the only σ-lattices among the classes $\mathbb{C}$ mentioned above.

The paper is organized as follows: In Section 2 estimators of the density contour clusters are given and the silhouette is defined. The connection of the silhouette to the Grenander estimator and to the estimator of Sager [20] are given in Section 2 also. Section 3 contains asymptotic results for the silhouette. Consistency results and rates of convergence in terms of L_1-distance are derived by means of empirical process theory. Section 4 contains some concluding remarks. Among others it is indicated there that the presented approach to density estimation can in principle also be applied to other (non-i.d.d.) situations such as nonparametric regression or spectral density estimation in time series analysis. All the proofs are given in Section 5.

2. The Silhouette

The following key equality holds for any density f:

$$f(x) = \int \mathbb{1}_{\Gamma(\lambda)}(x)\, d\lambda \qquad \forall x \in \mathbf{R}, \tag{2.1}$$

where $\mathbb{1}_{\Gamma(\lambda)}$ denotes the indicator function of $\Gamma(\lambda)$. A natural way to construct an estimator for f is to plug in the estimators of $\Gamma(\lambda)$ into (2.1). The silhouette is of this form (cf. (2.4) below). This idea is already inherent in the papers of Müller and Sawitzki [10, 11] (although not mentioned explicitly) from where the notion "silhouette" is taken.

Empirical Generalized λ-Clusters

Empirical generalized λ-clusters are used as estimators of $\Gamma(\lambda)$. They are defined as follows: Let $X_1, X_2, \ldots$ be i.i.d. random vectors in $\mathbf{R}^d$ with distribution F. Let F_n be the empirical distribution of the first n observations and let *Leb* denote Lebesgue measure.

Definition. Any set $\Gamma_{n,\mathbb{C}}(\lambda) \in \mathbb{C}$ such that

$$(F_n - \lambda \operatorname{Leb})(\Gamma_{n,\mathbb{C}}(\lambda)) = \sup_{C \in \mathbb{C}} (F_n - \lambda \operatorname{Leb})(C) \tag{2.2}$$

is called an *empirical generalized λ-cluster in* $\mathbb{C}$.

By definition $\Gamma_{n,\mathbb{C}}(\lambda)$ is a minimum volume set in $\mathbb{C}$ at level $\alpha = F_n(\Gamma_{n,\mathbb{C}}(\lambda))$. Since the sets $\Gamma_{n,\mathbb{C}}(\lambda)$ need not be connected, they are called *generalized* clusters. Hartigan [9] used the notion λ-cluster for maximal connected components of $\Gamma(\lambda)$. If F has Lebesgue density f then the

set $\Gamma(\lambda)$ maximizes $F-\lambda\,\mathrm{Leb}$ which is the theoretical counterpart of $F_n-\lambda\,\mathrm{Leb}$, i.e.,

$$(F-\lambda\,\mathrm{Leb})(\Gamma(\lambda))=\sup\{(F-\lambda\,\mathrm{Leb}(C),\ C\,\text{measurable}\}. \tag{2.3}$$

This motivates the definition of $\Gamma_{n,\mathbb{C}}(\lambda)$. Since F_n is a discrete and Leb is a continuous measure the supremum of $F_n-\lambda\,\mathrm{Leb}$ over all measurable sets equals one. Hence, in order to obtain a reasonable estimator of $\Gamma(\lambda)$ as a maximizer of $F_n-\lambda\,\mathrm{Leb}$ one has to restrict the maximization to certain subclasses $\mathbb{C}$ of subsets of $\mathbf{R}^d$.

In the one-dimensional case with $\mathbb{C}=\mathscr{I}_k$, where $\mathscr{I}_k$ denotes the class of all unions of at most k closed intervals, Müller and Sawitzki [11] proved the consistency of $\Gamma_{n,\mathbb{C}}(\lambda)$. Nolan [12] studied the sets $\Gamma_{n,\mathbb{C}}(\lambda)$ for $\mathbb{C}=\mathscr{E}^d$. She derived the asymptotic distribution of the corresponding finite-dimensional parameters. Under general conditions on $\mathbb{C}$ (analogous to those considered in the present paper) empirical generalized λ-clusters are studied in Polonik [13, 14].

As a function of λ the maximal value in (2.3), i.e., $E(\lambda)=(F-\lambda\,\mathrm{Leb})(\Gamma(\lambda))$, is called the *excess mass functional*. This notion has been introduced by Müller and Sawitzki [10]. The maximal value in (2.2), which is called the *empirical excess mass over* $\mathbb{C}$, serves as an estimator of the excess mass functional and can, for example, be used for constructing tests for multimodality (see Müller and Sawitzki [10, 11], Hartigan [8], Nolan [12], Polonik [13, 14]).

We assume below that

(A1) Empirical generalized λ-cluster in $\mathbb{C}$ exist for all $\lambda\geqslant 0$ and $\varnothing\in\mathbb{C}$.

(A1) is fulfilled for all classes of $\mathbb{C}$ mentioned above. For the class of convex sets this can be seen easily, because the supremum in (2.2) can be reduced to a maximum over convex polygons with vertices in the observations.

The sets $\Gamma_{n,\mathbb{C}}(\lambda)$ need not be uniquely determined. Nevertheless, for each choice of $\Gamma_{n,\mathbb{C}}(\lambda)$ a version of the *silhouette* can be defined as

$$f_{n,\mathbb{C}}(x)=\int \mathbb{1}_{\Gamma_{n,\mathbb{C}}(\lambda)}(x)\,d\lambda \qquad \forall x\in\mathbf{R}. \tag{2.4}$$

Müller and Sawitzki [11] implicitly used the silhouette in the one-dimensional case with $\mathbb{C}=\mathscr{I}_k$, $k=1, 2, \ldots$, to draw bootstrap samples out of the corresponding distribution. They proposed to use the silhouette as a data analytic tool.

To essentially overcome the non-uniqueness of the sets $\Gamma_{n,\mathbb{C}}(\lambda)$ we choose them such that

(A2) the function $\lambda \to \Gamma_{n,\mathbb{C}}(\lambda)$, $\lambda \geq 0$, is piecewise constant with at most n jumps. The jump points are denoted by $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_{k_n} = \lambda_{n,\max}$, $k_n \leq n$.

(A3) for every fixed $\lambda \geq 0$ the set $\Gamma_{n,\mathbb{C}}(\lambda)$ has minimal Lebesgue measure among all empirical generalized λ-clusters.

(A4) $\Gamma_{n,\mathbb{C}}(\lambda) = \varnothing$ for $\lambda \geq \lambda_{n,\max}$.

If (A1) holds, one can always find sets $\Gamma_{n,\mathbb{C}}(\lambda)$, $\lambda \geq 0$, such that (A2) and (A3) are satisfied (cf. proof of Proposition 2.2, Polonik [15]). Assumptions (A2) and (A3) do not affect the results given below, which hold for any choice of sets $\Gamma_{n,\mathbb{C}}(\lambda)$. (A4) actually is a definition. It is required in order to avoid that empirical λ-clusters have Lebesgue measure zero but have positive F_n-measure. Since F is assumed to be dominated by Lebesgue measure this seems to be a natural assumption. Under (A1)–(A4) the silhouette can be written as

$$f_{n,\mathbb{C}}(x) = \sum_{j=0}^{k_n - 1} (\lambda_{j+1} - \lambda_j) \mathbb{1}_{\Gamma_{n,\mathbb{C}}(\lambda_j)}(x). \tag{2.5}$$

If in addition the sets $\Gamma_{n,\mathbb{C}}(\lambda_j)$, $j = 0, \ldots, k_n$, are monotonically decreasing, i.e., $\Gamma_{n,\mathbb{C}}(\lambda_{j+1}) \subset \Gamma_{n,\mathbb{C}}(\lambda_j)$, then $f_{n,\mathbb{C}}$ can be visualized as putting the slices $\Gamma_{n,\mathbb{C}}(\lambda_j) \times [\lambda_j, \lambda_{j+1}]$ one on top of the other. Unfortunately, the monotonicity of the empirical λ-clusters need not hold such that $f_{n,\mathbb{C}}$ *does not lie in the model class* $\mathscr{F}_{\mathbb{C}}$ *in general.* But if $\mathbb{C}$ is a σ-lattice, then $f_{n,\mathbb{C}} \in \mathscr{F}_{\mathbb{C}}$ is always fulfilled (as, for example, for the Grenander estimator; see below). (A1)–(A4) are supposed to hold in all of what follows.

PROPOSITION 2.1. *If* $\lambda_{n,\max} > 0$, *then*

$$\int f_{n,\mathbb{C}}(x)\, dx = F_n(\Gamma_{n,\mathbb{C}}(0)).$$

Because of Proposition 2.1 we call $\mathbb{C}$ a *normalizing* class if $F_n(\Gamma_{n,\mathbb{C}}(0)) = 1$. All classes mentioned above are normalizing classes.

Connections to the Grenander Estimator and to the Estimator of Sager

It has been shown by Grenander [6] that the maximum likelihood estimator $\hat{f}_n$ of f in $\mathscr{F}_{\mathbb{M}_0}$ is given by the left-continuous slope of the smallest concave majorant of F_n. The symbol "F_n" is used for the empirical distribution function as well as for the empirical measure. To show that $f_{n,\mathbb{M}_0}$ equals $\hat{f}_n$ consider $U_n(\lambda) = \inf\{t \geq 0 : F_n(t) - \lambda t \text{ is maximal}\}$. U_n has the property that $\hat{f}_n(x) \leq \lambda \Leftrightarrow U_n(\lambda) \leq x$. (This fact is used in Groeneboom [7].)

Furthermore, we have $\Gamma_{n,\mathbb{M}_0}(\lambda) = [0, t_\lambda]$, where $t_\lambda \in \arg\max_{t \geq 0}\{F_n([0,t]) - \lambda \operatorname{Leb}([0,t])\} = \operatorname{argmax}_{t \geq 0}\{F_n(t) - \lambda t\}$. By (A3) this implies that $t_\lambda = U_n(\lambda)$. Using the monotonicity of the sets $\Gamma_{n,\mathbb{M}_0}(\lambda)$ we obtain

$$f_{n,\mathbb{M}_0}(x) \leq \lambda \Leftrightarrow x \notin \Gamma_{n,\mathbb{M}_0}(\lambda) \Leftrightarrow t_\lambda \leq x \Leftrightarrow U_n(\lambda) \leq x,$$

which shows the assertion. The following generalization of these arguments shows that if the sets $\Gamma_{n,\mathbb{C}}(\lambda_j)$, $j = 0, \ldots, k_n$, are nested then the silhouette is of the same structure as the estimator of Sager [20] (cf. Introduction). Let $C_n(\alpha)$ denote a minimum volume set in $\mathbb{C}$ at level α and let $\hat{F}_n$ be defined as follows: $\hat{F}_n(x) = F_n(C_n(j/n))$ for $x \in [\operatorname{Leb}(C_n(j/n)), \operatorname{Leb}(C_n((j+1)/n)))$, $j = 1, \ldots, n-1$ and $\hat{F}_n(x) = 0$ for $x < \operatorname{Leb}(C_n(1/n))$ and $\hat{F}_n(x) = F_n(C_n(j/n))$ for $x \geq \operatorname{Leb}(C_n(1))$. Note that $\hat{F}_n$ is a step function with $\hat{F}_n = F_n$ for $\mathbb{C} = \mathbb{M}_0$. As in the case of the Grenander estimator the smallest concave majorant of $\hat{F}_n$ gives the empirical generalized λ-clusters and the values λ_j of (2.5) as the slopes of the concave majorant. If the sets $\Gamma_{n,\mathbb{C}}(\lambda_j)$ are monotone then the values λ_j are the different values of the silhouette. Hence, in this case the silhouette is of the same type as the estimator of Sager [20] (the transfer density is estimated by the slope of the concave minorant of $\hat{F}_n$).

3. Asymptotic Results

In this section consistency results and rates of convergence for $f_{n,\mathbb{C}}$ are given in terms of L_1-distance denoted by $\|f - g\|_1 = \int |f(x) - g(x)|\, dx$. Let $(\Omega, \mathcal{A}, P)$ denote the underlying probability space. In order to avoid measurability considerations we define for any function $f: \Omega \to \mathbf{R}$ the *measurable cover function* f^* as the (pointwise) smallest measurable function from Ω to $\mathbf{R}$ lying everywhere above f, i.e., f^* measurable, $f^* \geq f$, and for all measurable g with $g \geq f$ we have $g \geq f^*$ almost surely. Of course, if f is measurable, then $f^* = f$. Let P^* denote the *outer probability*. Then for any $\alpha > 0$ we have $P^*(f > \alpha) = P(f^* > \alpha)$. See, for example, Dudley [4] for more details. We need the following definition.

Definition. $\mathbb{C}$ is called a Glivenko–Cantelli (GC)-class for F, or a GC(F)-class, if with probability 1,

$$(\sup_{C \in \mathbb{C}} |F_n(C) - F(C)|)^* \to 0 \qquad \text{as} \quad n \to \infty.$$

The classes $\mathbb{M}_0$, $\mathcal{I}_k$, $\mathcal{B}^d$, $\mathcal{E}^d$ are GC(F)-classes for all F. $\mathcal{C}^d$ is a GC(F)-class if F has a bounded Lebesgue density (see Eddy and Hartigan [5]). Moreover, all classes which are k-constructible from a GC(F)-class again

are GC(F)-classes. A class $\mathbb{C}$ in a measurable space $(\mathcal{X}, \mathcal{A})$ is called *k-constructible* from a class $\mathbb{D}$, if there exists a function φ from $\mathbb{D}^k$ to $\mathcal{A}$ constructed from $\cap$, $\cup$, ^{c} such that $\mathbb{C} \subset \varphi(\mathbb{D}^k)$ (this notion has been used by Alexander [1]).

THEOREM 3.1 (Consistency). *Suppose that $\mathbb{C}$ is a normalizing class. If $f \in \mathcal{F}_{\mathbb{C}}$, then there exists a real-valued (non-random) function $A = A(\eta, L)$, depending on f, with $A(\eta, L) \to 0$ as $\eta \to 0$ and $L \to \infty$, such that for all $L \geq \eta > 0$ we have*

$$\|f_{n,\mathbb{C}} - f\|_1 \leq 2\eta^{-1} \int_\eta^L [(F_n - F)(\Gamma_{n,\mathbb{C}}(\lambda)) - (F_n - F)(\Gamma(\lambda))]\, d\lambda + A(\eta, L). \tag{3.1}$$

Hence, if, in addition, $\mathbb{C}$ is a GC(F)*-class then we have with probability* 1 *that*

$$\|f_{n,\mathbb{C}} - f\|_1^* \to 0 \qquad \text{as} \quad n \to \infty.$$

Remark. Essentially Theorem 3.1 says that the GC(F)-property of $\mathbb{C}$ entails consistency of the silhouette. This also is the case for maximum likelihood estimators of densities measurable with respect to σ-lattices and has been noted by Sager [19] as a remarkable property.

Rates of Convergence

First we give the rates of convergence for Vapnik–Cervonenkis (VC)-classes $\mathbb{C}$. VC-classes are defined through a combinatorical property as follows: Let D be a finite set in $\mathbf{R}^d$. A class $\mathbb{C}$ is said to *shatter* D iff every $B \subset D$ is of the form $C \cap D$ for some $C \in \mathbb{C}$. If there exists a number $k \in \mathbf{N}$, such that $\mathbb{C}$ shatters no set which consists of k elements, then $\mathbb{C}$ is called a VC-class and the minimal k with that property is called the *index* of $\mathbb{C}$. Examples for VC-classes are $\mathbb{M}_0$, $\mathcal{I}_k$, the classes of all halfplanes, $\mathcal{B}^d$ and $\mathcal{E}^d$.

For the proofs of the theorems below we use results of Alexander [1] about the behaviour of set- and function-indexed empirical processes. We also use some of his terminology. A class $\mathbb{C}$ is called (v, k)-*constructible VC-class*, if $\mathbb{C}$ is k-constructible from a VC-class $\mathbb{D}$ whose index is smaller than or equal to v. A class $\mathbb{C}$ is called *n-deviation measurable*, if the corresponding $\mathbb{C}$-indexed empirical process satisfies a certain measurability condition. For the definition we refer to Alexander [1]. All VC-classes mentioned above are n-deviation measurable.

The rates of convergence of the silhouette depend on the tail behaviour of the underlying density f. With $\varphi(\lambda) = \mathrm{Leb}\{\Gamma(\lambda)\}$ the tail behaviour will here be measured in terms of the function

$$\Psi(\eta) = \Psi_f(\eta) = \int_0^\eta \varphi(\lambda)\, d\lambda.$$

DEFINITION. A level $\lambda > 0$ is called a critical level (of f), if φ is not differentiable at λ.

A level λ is critical if $\mathrm{Leb}\{x : f(x) = \lambda\} > 0$. In this case φ is not even continuous at λ and there exist sets Γ with $\mathrm{Leb}(\Gamma(\lambda)\, \Delta \Gamma) > 0$ also maximizing $F - \lambda\,\mathrm{Leb}$, where Δ denotes symmetric difference. If $\Gamma \in \mathbb{C}$ then $\Gamma_{n,\mathbb{C}}(\lambda)$ is not a consistent estimator of $\Gamma(\lambda)$ (cf. Polonik [14]). In view of (5.6) this in part explains the notion "critical."

THEOREM 3.2. *Let $\mathbb{C}$ be a normalizing n-deviation measurable (v, k)-constructible VC-class. Suppose that $\sup_x f(x) < \infty$ and that the number of critical levels of f is finite. If $f \in \mathcal{F}_{\mathbb{C}}$, then*

$$(\|f_{n,\mathbb{C}} - f\|_1)^* = O_P(\Psi(n^{-1/3}(\log n)^{1/3})) \qquad \text{as} \quad n \to \infty.$$

Remarks 3.3. (i) If the support of f has finite Lebesgue measure, then $\Psi(\eta) = O(\eta)$ as $\eta \to 0$ and the L_1-rate of the silhouette given in Theorem 3.2 is $n^{-1/3}(\log n)^{1/3}$. Otherwise, $\Psi(\eta)$ tends to zero (as $\eta \to 0$) slower than $O(\eta)$. This leads to slower rates of convergence of $f_{n,\mathbb{C}}$. For example, for the normal distribution in $\mathbf{R}^d$ one has $\Psi(\eta) = O(\eta(\log 1/\eta)^{d/2})$.

(ii) Since $\mathbb{M}_0$ is a VC-class, Theorem 3.2 also gives a rate of convergence for the Grenander estimator: If f has bounded support, then this rate is $n^{-1/3}(\log n)^{1/3}$. Groeneboom [7] showed that $n^{-1/3}$ is the exact rate under some smoothness conditions. Hence, we derived the exact rate up to a log-term using only the fact that the corresponding class $\mathbb{M}_0$ is a VC-class.

In the following theorem conditions on $\mathbb{C}$ are given in terms of metric entropy with inclusion of $\mathbb{C}$ with respect to F which is defined as follows. Let

$$\begin{aligned} N_I(\varepsilon, \mathbb{C}, F) := \inf\{ m \in \mathbf{N} : \exists C_1, \ldots, C_m \text{ measurable,} \\ \text{such that for every } C \in \mathbb{C} \text{ there exist} \\ i, j \in \{1, \ldots, m\} \text{ with } C_i \subset C \subset C_j \text{ and } F(C_j \backslash C_i) < \varepsilon \}, \end{aligned}$$

then $\log N_I(\varepsilon, \mathbb{C}, F)$ is called *metric entropy with inclusion of $\mathbb{C}$ with respect to F.*

THEOREM 3.4. *Let $\mathbb{C}$ be a normalizing class such that there exist constants A, $r>0$ with*

$$\log N_I(\varepsilon, \mathbb{C}, F) \leqslant A\varepsilon^{-r} \qquad \forall \varepsilon > 0. \tag{3.3}$$

Suppose that $\sup_x f(x) < \infty$ and that f has at most finitely many critical levels and let

$$\alpha_n = \begin{cases} n^{-1/(3+r)}, & r<1, \\ n^{-1/4}\log(n), & r=1, \\ n^{-1/2(r+1)}, & r>1. \end{cases}$$

Then $(\|f_{n,\mathbb{C}} - f\|_1)^* = O_P(\Psi(\alpha_n))$ as $n \to \infty$.

EXAMPLES. Let $\mathbb{C} = \mathscr{C}^d$, $d \geqslant 2$. If the support of f is compact and $\sup_x f(x) < \infty$, then (3.3) is satisfied with $r = (d-1)/2$ (see, e.g., Dudley [4]). Hence Theorem 3.4 gives in this case

$$\|f_{n,\mathscr{C}^d} - f\|_1^* = O_{P^*}\begin{pmatrix} n^{-2/7}, & d=2 \\ n^{-1/4}\log n, & d=3 \\ n^{-1/(d+1)}, & d \geqslant 4 \end{pmatrix}.$$

If the support of f is not compact, then (3.3) still holds with $r=(d-1)/2$ under mild conditions on the tail behaviour of f (see Polonik [13]). Therefore, up to an additional log-term (cf. Remark 3.3(i)) the given rates also hold for a normal distribution.

The Case of an Underlying Uniform Distribution

Theorem 3.2 and Theorem 3.4 can also be applied to an underlying uniform distribution. However, in this case they lead to rates which are far from optimal. For example, it is well know that the rate of convergence of the Grenander estimator for a uniform distribution is $O_P(n^{-1/2})$. However, Theorem 3.2 gives an upper bound of $n^{-1/3}(\log n)^{1/3}$. But we are able to re-derive the correct rate $n^{-1/2}$ up to a log-term (see Corollary 3.6 below).

THEOREM 3.5. *Let F be a uniform distribution on a set S with $\mathrm{Leb}(S) < \infty$. Let $\{\beta_n\}$ be a sequence of real numbers converging to zero as $n \to \infty$. Suppose that $\|F_n - F\|_{\mathbb{C}} = O_P(\beta_n)$ as $n \to \infty$. If $\mathbb{C}$ is a normalizing class and $S \in \mathbb{C}$ then*

$$(\|f_{n,\mathbb{C}} - f\|_1)^* = O_P(\beta_n \log(1/\beta_n)) \qquad \text{as} \quad n \to \infty.$$

COROLLARY 3.6. *Let F and $\mathbb{C}$ be as in Theorem* 3.5. *If in addition $\mathbb{C}$ satisfies the entropy condition* (3.3) *then*

$$(\|f_{n,\mathbb{C}} - f\|_1)^* = O_P(\alpha_n) \qquad \text{as} \quad n \to \infty, \tag{3.4}$$

where

$$\alpha_n = \begin{cases} n^{-1/2} \log n, & r < 1 \\ n^{-1/2} (\log n)^2, & r = 1 \\ n^{-1/(r+1)} \log n, & r > 1. \end{cases}$$

For n-deviation measurable (v, m)-constructible VC-classes (3.4) *holds with* $\alpha_n = n^{-1/2} \log n$.

4. CONCLUDING REMARKS

• *Computation.* In the one-dimensional case there exist fast algorithms of Müller and Sawitzki [11] for calculating empirical generalized λ-clusters and the corresponding silhouette for $\mathbb{C} = \mathcal{I}_k$, $k = 1, 2, 3, \ldots$. For $\mathbb{C} = \mathcal{C}^2$ Hartigan [8] gave an algorithm for directly calculating $\Gamma_{n,\mathcal{C}^2}(\lambda)$ for a fixed $\lambda \geq 0$. Since empirical generalized λ-clusters are minimum volume sets (at *random* levels α), it is possible to use algorithms for calculating minimum volume sets in oder to calculate empirical generalized λ-clusters. One first has to calculate all minimum volume sets, i.e., at *all* levels $\alpha = k/n$, $k = 1, \ldots, n$. Then, in a second step, it is easy to calculate all empirical generalized λ-clusters in $\mathcal{E}^d$. This is done by Nolan [12] for $\mathbb{C} = \mathcal{E}^d$. Recently we developed algorithms for calculating all minimum volume sets in $\mathcal{C}^2$ and in the class of unions of two disjoint convex sets in $\mathbf{R}^2$. These algorithms use ideas similar to Hartigan [8].

• *What happens if $f \notin \mathcal{F}_{\mathbb{C}}$?* Let $\Gamma_{\mathbb{C}}(\lambda)$ be a set of maximizing the signed measure $F - \lambda \operatorname{Leb}$ over the class $\mathbb{C}$, i.e., $(F - \lambda \operatorname{Leb})(\Gamma_{\mathbb{C}}(\lambda)) = \sup_{C \in \mathbb{C}} (F - \lambda \operatorname{Leb})(C)$. Every set $\Gamma_{\mathbb{C}}(\lambda)$ is called a *generalized λ-cluster.* Suppose that $\Gamma_{\mathbb{C}}(\lambda)$ exists for each $\lambda > 0$ and that it is unique (up to F-nullsets). It is shown in Polonik [13] that if $\mathbb{C}$ is normalizing, then $(\|f_{n,\mathbb{C}} - f_{\mathbb{C}}\|_1)^*$ converges to zero with probability 1, where $f_{\mathbb{C}}(x) = \int \mathbb{1}_{\Gamma_{\mathbb{C}}(\lambda)}(x)\, d\lambda \ \forall x \in \mathbf{R}$. Here $\mathbb{C}$ has to satisfy some additional assumptions (which are satisfied for the standard class $\mathcal{I}_1$, $\mathcal{B}^d$, $\mathcal{E}^d$, and $\mathcal{C}^d$). In Polonik [13] it is show under additional assumptions on $\mathbb{C}$, that the integral of $f_{\mathbb{C}}$ over $\mathbf{R}^d$ equals $F(\Gamma_{\mathbb{C}}(0))$. For standard classes $\mathbb{C}$ we have $F(\Gamma_{\mathbb{C}}(0)) = 1$.

Even if $\mathbb{C}$ consists of connected sets $f_{\mathbb{C}}$ may be multimodal if the underlying density is multimodal. For example, assume that f is a (smooth)

bimodal density on the real line; i.e., there exists a level $\lambda_0 \geq 0$ such that $\Gamma(\lambda) \in \mathcal{I}_1$ for all $\lambda < \lambda_0$ and $\Gamma(\lambda) \in \mathcal{I}_2 \setminus \mathcal{I}_1$ for all $\lambda \geq \lambda_0$. In addition assume that f has no flat parts. Then $f_{\mathbb{C}}$ is unimodal if one mode dominates the other measured by the excess mass. More precisely, let x_1 and x_2 denote the two modes (local maxima) of f. If for all $\lambda \geq \lambda_0$ the set $\Gamma_{\mathcal{I}_1}(\lambda)$ contains one fixed mode, then $f_{\mathcal{I}_1}$ is unimodal. Otherwise $f_{\mathcal{I}_1}$ is bimodal. This indicates, that a strongly bimodal silhouette calculated with intervals indicates that the underlying distribution actually is bimodal.

• *Extensions to non-i.i.d. situations.* The presented approach to estimating a density f on $\mathbf{R}^d$ can be extended to non-i.d.d. situations. For example, for estimating a spectral density f of a stationary time series consider the periodogram ordinates at the Fourier frequencies $\alpha_j = 2\pi j/n$, $j = 1, \ldots, n$ as observations. Replace the distribution and the empirical distribution in (2.2) and (2.3) by the spectral distribution $F(C) = \int_C f(\alpha)\, d\alpha$, $C \subset [0, 2\pi]$, and the empirical spectral distribution $F_n(C) = 2\pi/n \sum_{j:\, \alpha_j \in C} Y_j$, respectively. Instead of the usual empirical process as above the empirical spectral process comes up. Estimators for contour clusters of the spectral density and for the spectral density itself can be defined analogously to the empirical generalized λ-clusters and the silhouette. Well-known properties of the empirical spectral distribution can be used to derive consistency results and rates of convergence for these estimators. For example, it is known that F_n converges to F almost surely uniformly over VC-classes of sets. This leads to consistency results of the estimators. Rates of convergence can be derived by means of the fact that under entropy conditions on $\mathbb{C}$ the process $C \to n^{1/2}(F_n - F)(C)$, $C \in \mathbb{C}$, converges to a continuous Gaussian process (see Dahlhaus, 1988). An analogous approach can be used in nonparametric regression. There set-indexed partial sum processes appear. (For asymptotic properties of set-indexed partial sum processes see Alexander and Pyke [2] and Bass and Pyle [3]).

• *Modifications of the silhouette.* As already mentioned, the silhouette in general does not lie in $\mathcal{F}_{\mathbb{C}}$. One way is to avoid this fact (which, however, also has advantages; see above) is to construct an "iterative" silhouette; start at $\lambda = 0$ and by increasing λ only consider sets in $\mathbb{C}$ lying inside the foregoing empirical generalized λ-cluster. (One can also start at a maximal value λ and by decreasing λ only consider sets in $\mathbb{C}$ containing the foregoing empirical λ-clusters). However, in this case we do not only have to deal with one fixed class of sets, but the class of sets under consideration changes with λ and depends on the data. Moreover, the maximizing set of $F - \lambda$ Leb at a fixed λ (over the class under consideration at this level λ) is no longer $\Gamma(\lambda)$, in general. Therefore mathematical theory becomes more difficult.

5. Proofs

Proof of Proposition 2.1. First assume that the empirical generalized λ-clusters are monotone for inclusion. Let λ_j, $j=0, ..., k_n$, be the (random) levels of (A2). Note that the assumption $\lambda_{n,\max}>0$ is equivalent to $k_n \geq 1$. For every $j \in 0, ..., k_n-1$ we define $\Delta_{n,\mathbb{C}}(j) := \Gamma_{n,\mathbb{C}}(\lambda_j) \setminus \Gamma_{n,\mathbb{C}}(\lambda_{j+1})$. The sets $\Delta_{n,\mathbb{C}}(j)$, $j=0, ..., k_n-1$, are disjoint and

$$\bigcup_{j=0}^{k_n-1} \Delta_{n,\mathbb{C}}(j) = \Gamma_{n,\mathbb{C}}(0) \setminus \Gamma_{n,\mathbb{C}}(\lambda_{k_n}).$$

Together with (2.5) it follows that $f_{n,\mathbb{C}}(x) = \lambda_{j+1}$ for all $x \in \Delta_{n,\mathbb{C}}(j)$. Hence

$$\int f_{n,\mathbb{C}}(x)\,dx = \sum_{j=0}^{k_n-1} \int_{\Delta_{n,\mathbb{C}}(j)} f_{n,\mathbb{C}}(x)\,dx = \sum_{j=0}^{k_n-1} \lambda_{j+1} \operatorname{Leb}(\Delta_{n,\mathbb{C}}(j)). \tag{5.1}$$

It follows from the definition of the empirical excess mass $E_{n,\mathbb{C}}(\lambda) = (F_n - \lambda \operatorname{Leb})(\Gamma_{n,\mathbb{C}}(\lambda))$ (cf. (2.2)) that for every $j \in \{0, ..., k_n-1\}$ an empirical generalized λ_j-cluster also is an empirical generalized λ_{j+1}-cluster (for details see Polonik [15, Proof of Proposition 2.2]). This means, that for every $j \in \{0, ..., k_n-1\}$ we have $F_n(\Gamma_{n,\mathbb{C}}(\lambda_{j+1})) - \lambda_{j+1} \operatorname{Leb}(\Gamma_{n,\mathbb{C}}(\lambda_{j+1})) = F_n(\Gamma_{n,\mathbb{C}}(\lambda_j)) - \lambda_{j+1} \operatorname{Leb}(\Gamma_{n,\mathbb{C}}(\lambda_j))$ and, hence,

$$\begin{aligned}\lambda_{j+1} &= [F_n(\Gamma_{n,\mathbb{C}}(\lambda_{j+1})) - F_n(\Gamma_{n,\mathbb{C}}(\lambda_j))] / [\operatorname{Leb}(\Gamma_{n,\mathbb{C}}(\lambda_{j+1})) - \operatorname{Leb}(\Gamma_{n,\mathbb{C}}(\lambda_j))]. \\ &= F_n(\Delta_{n,\mathbb{C}}(j)) / \operatorname{Leb}(\Delta_{n,\mathbb{C}}(j)).\end{aligned}$$

This, together with (5.1) and (A4), gives the assertion. The case without the monotonicity assumption can be reduced to the special case just proven as follows. Let $S_{n,\mathbb{C}}(\lambda_j) := \Gamma_{n,\mathbb{C}}(\lambda_{j-1}) \times [\lambda_j, \lambda_{j-1}]$, $j=1, ..., k_n$. No matter if the sets $\Gamma_{n,\mathbb{C}}(\lambda_j)$ are monotone or not the volume of $S_{n,\mathbb{C}} = \bigcup_{j=1}^{k_n-1} S_{n,\mathbb{C}}(\lambda_j) \subset \mathbf{R}^{d+1}$ equals the L_1-norm of $f_{n,\mathbb{C}}(x)$. The volume of $S_{n,\mathbb{C}}$ does not change if the sets $\Gamma_{n,\mathbb{C}}(\lambda_j)$ are replaced by their so-called "Schwarz symmetrizations" $\tilde{\Gamma}_{n,\mathbb{C}}(\lambda)$. They are defined as balls with midpoint zero which have the same Lebesgue measure as $\Gamma_{n,\mathbb{C}}(\lambda)$. The sets $\tilde{\Gamma}_{n,\mathbb{C}}(\lambda)$ are monotonically decreasing (in λ) for inclusion, because the Lebesgue measures of the empirical λ-clusters are decreasing in λ. (This property follows from the fact that by definition the empirical excess mass $E_{n,\mathbb{C}}(\lambda)$ is a convex, piecewise linear, decreasing function of λ with right-hand derivative $-\operatorname{Leb}(\Gamma_{n,\mathbb{C}}(\lambda))$). Hence, replacing the sets $\Gamma_{n,\mathbb{C}}(\lambda)$ by $\tilde{\Gamma}_{n,\mathbb{C}}(\lambda)$ and putting the resulting symmetrized slices $\tilde{\Gamma}_{n,\mathbb{C}}(\lambda_{j-1,n}) \times [\lambda_j, \lambda_{j-1}]$, $j=1, ..., k_n$, one on the top of the other gives us a function $\tilde{f}_{n,\mathbb{C}}$ which has the same L_1-norm as $f_{n,\mathbb{C}}$ and whose L_1-norm can be calculated as in the "case of monotonicity" treated above. ∎

Proof of Theorem 3.1. For any measurable set $A \subset [0, \infty)$ define

$$f_{n,\mathbb{C}}(x, A) := \int_A \mathbb{1}_{\Gamma_{n,\mathbb{C}}(\lambda)}(x)\, d\lambda.$$

Analogously define $f(x, A)$ with $\Gamma(\lambda)$ instead of $\Gamma_{n,\mathbb{C}}(\lambda)$. Let "$\Delta$" denote the symmetric difference and for $0 < \eta < L$; let $A_{\eta, L} = [\eta, L]$. The following inequality is proven below:

$$\|f_{n,\mathbb{C}} - f\|_1 \leq 2 \int_\eta^L \mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)\, \Delta \Gamma(\lambda))\, d\lambda + 2 \int f(x, (A_{\eta, L})^c)\, dx. \tag{5.2}$$

To estimate the first integral on the right-hand side of (5.2) we use the following fact: if $\Gamma(\lambda) \in \mathbb{C}$ then we have for every $\eta > 0$

$$\begin{aligned}\mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)\, \Delta \Gamma(\lambda)) \leq\ & \mathrm{Leb}\{x : |f(x) - \lambda| < \eta\} \\ & + \eta^{-1}[(F_n - F)(\Gamma_{n,\mathbb{C}}(\lambda)) - (F_n - F)(\Gamma(\lambda))]. \end{aligned} \tag{5.3}$$

This inequality has already been used in Polonik [13, 14]. For completeness we give the proof of (5.3) below. Put (5.3) in (5.2) to obtain

$$\begin{aligned}&\|f_{n,\mathbb{C}} - f\|_1 \\ &\quad \leq 2\eta^{-1} \int_\eta^L [(F_n - F)(\Gamma_{n,\mathbb{C}}(\lambda)) - (F_n - F)(\Gamma(\lambda))]\, d\lambda + A(\eta, L) \end{aligned} \tag{5.4}$$

with

$$A(\eta, L) := 2 \int f(x, (A_{\eta, L})^c)\, dx + 2 \int_\eta^L \mathrm{Leb}\{x : |f(x) - \lambda| < \eta\}\, d\lambda.$$

Since $\mathbb{C}$ is a GC(F)-class the measurable cover of the first term on the right-hand side of (5.4) converges to zero almost surely for any fixed η and L. It remains to show that $A(\eta, L) \to 0$ as $\eta \to 0$ and $L \to \infty$. The first integral in the definition of $A(\eta, L)$ can be written as $2 \int_M^\infty \varphi(\lambda)\, d\lambda + 2 \int_0^\eta \varphi(\lambda)\, d\lambda$. Both of these integrals can be made arbitrarily small be choosing L large enough and η small enough, respectively. (Note that $\int_L^\infty \varphi(\lambda)\, d\lambda = \int (f(x) - L)^+\, dx$ and $\int_0^\eta \varphi(\lambda)\, d\lambda = \int (\eta \wedge f(x))\, dx$). As for the second integral note that for η small enough

$$\begin{aligned}&\mathrm{Leb}\{x : |f(x) - \lambda| < \eta\} \\ &\quad = \varphi(\lambda - \eta) - \varphi(\lambda + \eta) - \mathrm{Leb}\{x : f(x) = \lambda - \eta\}. \end{aligned} \tag{5.5}$$

There exist at most countable many levels μ with $\text{Leb}\{x : f(x) = \mu\} \neq 0$. Hence, it follows that

$$\int_\eta^L \text{Leb}\{x : |f(x) - \lambda| < \eta\}\, d\lambda = \int_\eta^L \varphi(\lambda - \eta) - \varphi(\lambda + \eta)\, d\lambda$$
$$= \int_0^{2\eta} \varphi(\lambda)\, d\lambda - \int_{L-\eta}^{L+\eta} \varphi(\lambda)\, d\lambda.$$

Since f is integrable the last two integrals converge to zero as $\eta \to 0$ and, therefore, $A(\eta, L) \to 0$ as $\eta \to 0$ and $L \to \infty$. ∎

Proof of Theorem 3.2 *and Theorem* 3.4. The idea of the proof is as follows. It is easy to see that

$$\|f_{n,\mathbb{C}} - f\|_1 \leqslant \int \text{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)\, \Delta \Gamma(\lambda))\, d\lambda \tag{5.6}$$

(equality holds in (5.6) if all the sets $\Gamma_{n,\mathbb{C}}(\lambda)$ are level sets of $f_{n,\mathbb{C}}$.) Here "Δ" again denotes symmetric difference. We show (see Lemma 5.1 below) that there exists a function $k_n(\lambda)$ such that for a "large enough" region $\Lambda_n \subset [0, \infty)$.

$$\sup_{\lambda \in \Lambda_n} [k_n(\lambda)\, \text{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)\, \Delta \Gamma(\lambda))] = O_{P^*}(\alpha_n),$$

where

$$\alpha_n = \begin{cases} n^{1/3}(\log n)^{1/3} & \text{if } \mathbb{C} \text{ is a } VC\text{-class} \\ \text{defined as in Theorem 3.4} & \text{if } \mathbb{C} \text{ satisfies (3.3)} \end{cases}$$

are the rates asserted in the theorems. If in addition $1/k_n$ is integrable over Λ_n, then it follows that

$$\|f_{n,\mathbb{C}} - f\|_1^* = O_P\left(\alpha_n \int_{\Lambda_n} k_n(\lambda)^{-1}\, d\lambda\right).$$

It will turn out that $\alpha_n \int_{\Lambda_n} k_n(\lambda)^{-1}\, d\lambda = \Psi(\alpha_n)$, such that the assertions of Theorem 3.2 and Theorem 3.4, respectively, follow.

Now we give the exact proof. To simplify the notation we assume that f has only one critical level $\lambda_0 > 0$. (The proof for the case of more than one critical level is completely analogous.) Let $M = \sup f(x) < \infty$. From (5.3) we get with $\Lambda = (\alpha_n, M)$ that

$$\|f_{n,\mathbb{C}} - f\|_1 \leqslant 2 \int_{\alpha_n}^M \text{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)\, \Delta \Gamma(\lambda)) + 2 \int_0^{\alpha_n} \varphi(\lambda)\, d\lambda. \tag{5.7}$$

The last integral in (5.7) is of the asserted order. As for the first integral on the right-hand side of (5.7) we have the following. First assume that $\lambda_0 < M$ and write $\int_{\alpha_n}^{M} \mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)\, \Delta\Gamma(\lambda))\, d\lambda = (\int_{\alpha_n}^{\lambda_0-\alpha_n} + \int_{\lambda_0-\alpha_n}^{\lambda_0+\alpha_n} + \int_{\lambda_0+\alpha_n}^{M}) \mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)\, \Delta\Gamma(\lambda))\, d\lambda$. Let these three integrals (in the given order) be denoted by I_1, I_2, and I_3, respectively. I_2 is of the order $O_P(\alpha_n)$, because for any fixed $\varepsilon > 0$ we have for large enough n (such that $\alpha_n < \varepsilon$) that

$$\begin{aligned}
&\sup_{\lambda \in (\lambda_0-\alpha_n,\, \lambda_0+\alpha_n)} \mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)\, \Delta\Gamma(\lambda)) \\
&\leqslant \sup_{\lambda \in (\lambda_0-\alpha_n,\, \lambda_0+\alpha_n)} [\mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)) + \mathrm{Leb}(\Gamma(\lambda))] \\
&\leqslant \mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda_0-\varepsilon)) + \mathrm{Leb}(\Gamma(\lambda_0-\varepsilon)) \\
&\leqslant 2\, \mathrm{Leb}(\Gamma(\lambda_0-\varepsilon)) + O_P(1) = O_P(1).
\end{aligned}$$

The second of these inequality follows from the monotonicity of the functions $\lambda \to \mathrm{Leb}(\Gamma(\lambda))$ and $\lambda \to \mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda))$ (Polonik [15, Proposition 2.2]; see also proof of Proposition 2.1 given above). The last inequality follows from the fact that $\Gamma_{n,\mathbb{C}}(\lambda)$ is a consistent estimator of $\Gamma(\lambda)$ for all λ that are no critical level (Polonik [13, 14]). In order to control I_1 and I_3 we need the following lemma (it is proved below). Let φ' denote the derivative of φ with respect to λ.

LEMMA 5.1. *Suppose that the assumptions of Theorem* 3.2 *and Theorem* 3.4, *respectively, hold. Let* α_n *be as above and let* $0 < a_n < b_n \leqslant \infty$ *be such that the interval* $(a_n - \alpha_n,\, b_n + \alpha_n]$ *contains no critical level. For* $\lambda \in (a_n, b_n)$ *let* $\xi_{\lambda,n}$ *be defined through the equation*

$$\mathrm{Leb}\{x : |f(x) - \lambda| < \alpha_n\} = \varphi(\lambda - \alpha_n) - \varphi(\lambda + \alpha_n) = -2\varphi'(\xi_{\lambda,n})\alpha_n.$$

Then we have with $h_n(\lambda) = [\,|\varphi'(\xi_{\lambda,n})| \vee 1\,]^{-1}$ *that as* $n \to \infty$

$$(\sup_{\lambda \in (a_n,\, b_n)} [h_n(\lambda)\, \mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)\, \Delta\Gamma(\lambda))])^* = O_P(\alpha_n).$$

COROLLARY. *In the situation of Lemma* 5.1 *we have as* $n \to \infty$

$$\int_a^b \mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)\, \Delta\Gamma(\lambda))\, d\lambda = O_{P^*}(\alpha_n) + O_{P^*}\left(\int_{a-\alpha_n}^{a+\alpha_n} (\lambda)\, d\lambda\right).$$

Proof.

$$\begin{aligned}
&\int_a^b \mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)\,\Delta\Gamma(\lambda))\,d\lambda \\
&\leqslant O_{P^*}(\alpha_n)\int_a^b [\,|\varphi'(\xi_{\lambda,n})| \vee 1\,]\,d\lambda \\
&\leqslant O_{P^*}(\alpha_n)\left[(b-a)+\int_a^b |\varphi'(\xi_{\lambda,n})|\,d\lambda\right] \\
&= O_{P^*}(\alpha_n)+O_{P^*}\left(\int_{a-\alpha_n}^{a+\alpha_n}\varphi(\lambda)\,d\lambda-\int_{b-\alpha_n}^{b+\alpha_n}\varphi(\lambda)\,d\lambda\right) \\
&\leqslant O_{P^*}(\alpha_n)+O_{P^*}\left(\int_{a\quad\alpha_n}^{a+\alpha_n}\varphi(\lambda)\,d\lambda\right).
\end{aligned}$$

The third line follows from definition of $\xi_{\lambda,n}$ and the last inequality holds since φ is decreasing. ∎

The above corollary implies $I_1 \leqslant O_{P^*}(\alpha_n)+O_{P^*}(\int_0^{2\alpha_n}\varphi(\lambda)\,d\lambda) = O_{P^*}(\Psi(\alpha_n))$ and $I_3 \leqslant O_{P^*}(\alpha_n)+O_{P^*}(\int_{\lambda_0}^{\lambda_0+2\alpha_n}\varphi(\lambda)\,d\lambda) = O_{P^*}(\alpha_n)$ and the assertion follows. If $\lambda_0 = M$ we split $\int_{\alpha_n}^{M}\mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)\,\Delta\Gamma(\lambda))\,d\lambda$ into two integrals extended over $(\alpha_n, \lambda_0-\alpha_n)$ and $(\lambda_0-\alpha_n, \lambda_0)$, respectively. Upper bounds for these two integrals can be obtained as above. ∎

Proof of Theorem 3.5 *and Corollary* 3.6. Let $M = 1/\mathrm{Leb}(S)$. By the definition of $\Gamma_{n,\mathbb{C}}(\lambda)$ we have $0 \leqslant (F_n - \lambda\,\mathrm{Leb})(\Gamma_{n,\mathbb{C}}(\lambda)) - (F_n - \lambda\,\mathrm{Leb})(S)$. It follows by writing $F_n = F + F_n - F$ that

$$(F-\lambda\,\mathrm{Leb})(S)-(F-\lambda\,\mathrm{Leb})(\Gamma_{n,\mathbb{C}}(\lambda)) \leqslant (F_n-F)(\Gamma_{n,\mathbb{C}}(\lambda))+(F_n-F)(S). \tag{5.8}$$

Since, furthermore, $(F - \lambda\,\mathrm{Leb})(S) - (F - \lambda\,\mathrm{Leb})(\Gamma_{n,\mathbb{C}}(\lambda)) = (M - \lambda)\,\mathrm{Leb}(S\backslash\Gamma_{n,\mathbb{C}}(\lambda)) + \lambda\,\mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)\backslash S) \geqslant (M-\lambda)\,\mathrm{Leb}(S\backslash\Gamma_{n,\mathbb{C}}(\lambda))$, we have

$$(M-\lambda)\,\mathrm{Leb}(S\backslash\Gamma_{n,\mathbb{C}}(\lambda)) \leqslant 2\sup_{C\in\mathbb{C}}|(F_n-F)(C)|. \tag{5.9}$$

By using the fact that $S = \Gamma(\lambda)$ for $\lambda < M$ it is easy to verify that

$$\|f_{n,\mathbb{C}} - f\|_1 \leqslant 2\int_0^M \mathrm{Leb}(S\backslash\Gamma_{n,\mathbb{C}}(\lambda))\,d\lambda.$$

Together with (5.9), it follows that

$$\|f_{n,\mathbb{C}} - f\|_1 \leqslant 2\left[\int_0^{M-\beta_n} + \int_{M-\beta_n}^{M}\right] \mathrm{Leb}(S \backslash \Gamma_{n,\mathbb{C}}(\lambda))\, d\lambda$$

$$= O_P(\beta_n) \int_0^{M-\beta_n} (M(S)-\lambda)^{-1}\, d\lambda + \int_{M-\beta_n}^{M} \mathrm{Leb}(S \backslash \Gamma_{n,\mathbb{C}}(\lambda))\, d\lambda$$

The first term in the last line is of the order $O_{P^*}(\beta_n)\, O(\log 1/\beta_n)$. The second term is of the order $O_{P^*}(\beta_n)$. This proves the first assertion of the theorem. Corollary 3.6 follows from the fact that $\|F_n - F\|_{\mathbb{C}} = O_{P^*}(n^{-1/2})$ for VC-classes $\mathbb{C}$ and that for classes $\mathbb{C}$ satisfying (3.3) one has (see Alexander [1])

$$\|F_n - F\|_{\mathbb{C}} = O_P\begin{pmatrix} n^{-1/2}, & r<1 \\ n^{-1/2}\log n, & r=1 \\ n^{-1/(r+1)}, & r>1 \end{pmatrix}. \quad \blacksquare$$

Proof of inequality (5.3). For any $\eta > 0$ we have

$$(F - \lambda\,\mathrm{Leb})(\Gamma(\lambda)) - (F - \lambda\,\mathrm{Leb})(C)$$

$$= (F - \lambda\,\mathrm{Leb})(\Gamma(\lambda) \backslash C) - (F - \lambda\,\mathrm{Leb})(C \backslash \Gamma(\lambda))$$

$$= \int_{\Gamma(\lambda)\, \Delta\, C} |f(x) - \lambda|\, dx$$

$$\geqslant \eta\, \mathrm{Leb}((\Gamma_{n,\mathbb{C}}(\lambda)\, \Delta \Gamma_{\mathbb{C}}(\lambda)) \cap \{x : |f(x) - \lambda| \geqslant \eta\}). \tag{5.10}$$

Since $\mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)\, \Delta \Gamma_{\mathbb{C}}(\lambda)) \leqslant \mathrm{Leb}(x : |f(x) - \lambda| < \eta) + \mathrm{Leb}((\Gamma_{n,\mathbb{C}}(\lambda)\, \Delta \Gamma_{\mathbb{C}}(\lambda)) \cap \{x : |f(x) - \lambda| \geqslant \eta\})$ the assertion easily follows from (5.10) and (5.8) with S replaced by $\Gamma(\lambda)$. $\blacksquare$

Proof of (5.2). For every given set $A \subset [0, \infty)$ it immediately follows from the definition of $f(x, A)$ that $f(x) = f(x, A) + f(x, A^c)$. The analogous decomposition holds for $f_{n,\mathbb{C}}$. Hence,

$$\|f_{n,\mathbb{C}} - f\|_1 \leqslant \|f_{n,\mathbb{C}}(\cdot, A) - f(\cdot, A)\|_1 + \int f_{n,\mathbb{C}}(x, A^c)\, dx$$

$$+ \int f(x, A^c)\, dx. \tag{5.11}$$

Using $\int f_{n,\mathbb{C}}(x)\, dx = 1$ (Proposition 2.1) we have for the second integral on the right-hand side of (5.11):

$$\int f_{n,\mathbb{C}}(x, A^c)\,dx = 1 - \int f_{n,\mathbb{C}}(x, A)\,dx$$

$$= \int f(x, A) - f_{n,\mathbb{C}}(x, A)\,dx + \int f(x, A^c)\,dx$$

$$\leqslant \|f_{n,\mathbb{C}}(\cdot, A) - f(\cdot, A)\|_1 + \int f(x, A^c)\,dx.$$

Together with (5.6), the assertion follows. ∎

Proof of Lemma 5.1. For any measure F and any integrable function g let $F(g) = \int g\,dF$. Choose $\eta = \alpha_n$ in (5.3). Then multiplication of (5.3) by $h_n(\lambda)$ leads by definition of $h_n(\lambda)$ to

$$\|g_n(\lambda)\|_1 \leqslant 2\alpha_n + \alpha_n^{-1}(F_n - F)(g_n(\lambda)), \tag{5.12}$$

where $g_n(\lambda) = h_n(\lambda)(\mathbb{1}_{I_{n,\mathbb{C}}(\lambda)\setminus I(\lambda)} - \mathbb{1}_{I(\lambda)\setminus I_{n,\mathbb{C}}(\lambda)})$. Let $\mathscr{G}_{\mathbb{C}} = \{r(\mathbb{1}_{C\setminus D} - \mathbb{1}_{D\setminus C}),$ $r \leqslant 1,\ C, D \in \mathbb{C}\}$. For $k > 0$ define

$$B_n = \{\exists g \in \mathscr{G}_{\mathbb{C}}, \text{ such that } \|g\|_1 > k\alpha_n$$
$$\text{and } 1 \leqslant 2k^{-1} + \alpha_n^{-1}\,|(F_n - F)(g)|/\|g\|_1\}.$$

The assertion of the theorem follows if we show that there exists a constant $k > 0$ such that $P^*(B_n) \to 0$ as $n \to \infty$. In order to prove this we show that there exists a constant $k > 4$ with

$$P^*\Big(\sup_{g \in \mathscr{G}_{\mathbb{C}}\,:\,\|g\|_1 > k\alpha_n} |v_n(g)|/\|g\|_1 > n^{1/2}\alpha_n/2\Big) \to 0 \qquad \text{as} \quad n \to \infty,$$

where $v_n = n^{1/2}(F_n - F)$. In the following the supremum is always extended over $g \in \mathscr{G}_{\mathbb{C}}$ which satisfy certain conditions. To shorten the notation we omit "$g \in \mathscr{G}_{\mathbb{C}}$." We have

$$P^*\Big(\sup_{\|g\|_1 > k\alpha_n} |v_n(g)|/\|g\|_1 > n^{1/2}\alpha_n/2\Big)$$

$$\leqslant \sum_{j=0}^{\infty} P^*\Big(\sup_{k2^j\alpha_n < \|g\|_1 \leqslant k2^{j+1}\alpha_n} |v_n(g)|/\|g\|_1 > n^{1/2}\alpha_n/2\Big)$$

$$\leqslant \sum_{j=0}^{\infty} P^*\Big(\sup_{\|g\|_1 \leqslant k2^{j+1}\alpha_n} |v_n(g)| > k2^{j-1}n^{1/2}\alpha_n^2\Big)$$

$$= \sum_{j=1}^{\infty} P^*\Big(\sup_{\|g\|_1 \leqslant k2^j\alpha_n} |v_n(g)| > k2^j n^{1/2}\alpha_n^2\Big) = \sum_{j=1}^{\infty} p_{n,j}.$$

To show that the last sum converges to zero as n tends to infinity we use results of Alexander [1]. For VC-classes $\mathbb{C}$ his results can be used directly. For the case that $\mathbb{C}$ satisfies (3.3) we need a minor extention (Theorem 5.2) of his results from the set-indexed to the function-indexed empirical process. For a class $\mathcal{G}$ of functions with $\mathcal{G} \subset L_p(F)$ let

$$N_{p,B}(\varepsilon, \mathcal{G}, F) := \inf\{m \in \mathbf{N} : \exists g_1, \ldots, g_m \text{ measurable,}$$
$$\text{such that for every } g \in \mathcal{G} \text{ there exist}$$
$$i, j \in \{1, \ldots, m\} \text{ with } g_i \leqslant g \leqslant g_j \text{ and } \|g_j - g_i\|_{p,F} < \varepsilon\},$$

where $\|\cdot\|_{p,F}$ denotes the L_p-norm with respect to F. Then $\log N_{p,B}(\varepsilon, \mathcal{G}, F)$ is called *metric entropy with bracking of* $\mathcal{G}$ *in* $L_p(F)$.

THEOREM 5.2 (Alexander). *Let* $\mathcal{G}$ *be a class of functions with* $0 \leqslant g \leqslant 1$. *Let* $L(x) = \log(x \vee e)$, $n \geqslant 1$, $\alpha \geqslant \sup_{g \in \mathcal{G}} \operatorname{var} \nu_n(g)$, *with* $\nu_n = n^{1/2}(F_n - F)$ *and define* $\Psi(L, n, \alpha) = L^2/2\alpha(1 + L/3n^{1/2}\alpha)$. *Suppose that*

$$\log N_{1,B}(\varepsilon, \mathcal{G}, F) \leqslant A\varepsilon^{-r} \qquad \forall \varepsilon > 0.$$

Then there exist constants $K_i = K_i(r, A)$, $i = 1, 2, 3$, *such that if* $L \leqslant n^{1/2}\alpha_n^2/32$ *and*

$$L \geqslant \begin{cases} K_1 \alpha^{(1-r)/2} & \text{if } r < 1 \\ K_2 L(n) & \text{if } r = 1 \\ K_3 n^{(r-1)/2(r+1)} & \forall r \end{cases}$$

then

$$P^*(\sup_{g \in \mathcal{G}} |\nu_n(g)| > L) \leqslant 5 \exp\{-1/2\Psi(L, n, \alpha)\}.$$

To prove this theorem one only has to replace δ_j by δ_j^2 and use L_1-bracketing functions instead of L_2-bracketing functions in the proof of Alexander [1, Corollary 2.4].

Let $M = \sup\{f(x)\} < \infty$. Define $\mathcal{G}_{\mathbb{C}}(j, n) = \{g \in \mathcal{G}_{\mathbb{C}} : \|g\|_{1,F} < Mk2^j\alpha_n\}$, then we have $p_{n,j} \leqslant P^*(\sup_{\mathcal{G}_{\mathbb{C}}(j,n)} |\nu_n(f)| > C2^j n^{1/2}\alpha_n^2)$. Note that for any $g \in \mathcal{G}_{\mathbb{C}}$ we have $\operatorname{var}(\nu_n(g)) \leqslant \|g\|_{2,F}^2 \leqslant \|g\|_{1,F} \leqslant M\|g\|_1$, so that $\sup_{g \in \mathcal{G}_{\mathbb{C}}(j,n)} \operatorname{var}(\nu_n(g)) \leqslant Mk2^j\alpha_n$. Now we apply Theorem 5.2 with $\alpha = Mk2^j\alpha_n$ and $L = k2^j n^{1/2}\alpha_n^2$. This is easy to verify that if $\log N_1(\varepsilon, \mathbb{C}, F) = \mathcal{O}(\varepsilon^{-r})$, then $\log N_{1,B}(\varepsilon, \mathcal{G}_{\mathbb{C}}, F)$ is of the same order. The conditions of Theorem 5.2 are satisfied with this choice of α and L. This follows by elementary calculations, which in addition show, that the conditions are satisfied for all $k > k_0$, with k_0 large enough, independent of n and j. Therefore k can be chosen to be bigger than 4 as required. It follows that

$$\sum_{j=1}^{\infty} p_{n,j} \leq 5 \sum_{j=1}^{\infty} \exp\{-1/2\,\Psi(k2^j n^{1/2}\alpha_n^2, n, MC2^j\alpha_n)\}$$

$$= 5 \sum_{j=1}^{\infty} \exp\{-(k2^j n\alpha_n^3)/(2M(1+\alpha_n/3M)\}.$$

Since $\alpha_n \to 0$ as $n \to \infty$ and $n\alpha_n^3 \geq \log n$ it follows that $\sum_{j=1}^{\infty} p_{n,j} \to 0$ as $n \to \infty$. ∎

ACKNOWLEDGMENTS

This paper is a revised version of a part of my thesis. I thank my supervisor Professor D. W. Müller for his interest and support during the time of writing my thesis. Furthermore, I thank the statistics group of the Universität Heidelberg, in particular W. Ehm, for valuable discussions, hints, and remarks concerning the subject.

REFERENCES

[1] ALEXANDER, K. S. (1987). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **15** 428–430.

[2] ALEXANDER, K. S., AND PYKE, R. (1986). A uniform central limit theorem for set-indexed partial-sum processes with finite variance. *Ann. Probab.* **14** 582–597.

[3] BASS, R. F., AND PYKE, R. (1984). Functional law of the iterated logarithm and uniform central limit theorem for partial-sum processes indexed by sets. *Ann. Probab.* **12** 13–34.

[4] DUDLEY, R. M. (1984). A course on empirical processes. In *École d'Été de Probabilités de Saint Flour XII-1982*, Lecture Notes in Math., Vol. 1097, pp. 1–142. Springer-Verlag, New York.

[5] EDDY, W. F., AND HARTIGAN, J. A. (1977). Uniform convergence of the empirical distribution function over convex sets. *Ann. Statist.* **5** 370–374.

[6] GRENANDER, U. (1956). On the theory of mortality measurement, Part II. *Skand. Akt.* **39** 125–153.

[7] GROENEBOOM, P. (1985). Estimating a monotone density. In *Proceedings, Berkeley Conference in Honor of Jerzy Neymann and Jack Kiefer*, Vol. II (L. LeCam and R. Olshen, Eds.). Wadsworth, Monterey, CA.

[8] HARTIGAN, J. A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82** 267–270.

[9] HARTIGAN, J. A. (1975). *Clustering Algorithms*. Wiley, New York/London.

[10] MÜLLER, D. W., AND SAWITZKI, G. (1987). Using excess mass estimates to investigate the modality of a distribution. Preprint 398, SFB 123, Universität Heidelberg.

[11] MÜLLER, D. W., AND SAWITZKI, G. (1991). Excess mass estimates and tests of multimodality. *J. Amer. Statist. Assoc.* **86** 738–746.

[12] NOLAN, D. (1991). The excess-mass ellipsoid. *J. Multivariate Anal.* **39** 348–371.

[13] POLONIK, W. (1992). *The Excess Mass Approach to Cluster Analysis and Related Estimation Procedures*. Dissertation, Universität Heidelberg.

[14] POLONIK, W. (1995). Measuring mass concentrations and estimating density contour clusters—An excess mass approach. *Ann. Stat.*, to appear.

[15] POLONIK, W. (1993). *Density Estimation Under Qualitative Assumptions in Higher Dimensions.* Beiträge zur Statistik Nr. 15, Institut für Angewandte Mathematik, Universität Heidelberg.

[16] ROBERTSON, T. (1967). On estimating a density measurable with respect to a σ-lattice. *Ann. Math. Statist.* **38** 482–493.

[17] ROBERTSON, T., WRIGHT, F. T., AND DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference.* Wiley, New York.

[18] SAGER, T. W. (1979). An iterative method for estimating a multivariate mode and isopleth. *J. Amer. Statist. Assoc.* **74** 329–339.

[19] SAGER, T. W. (1982). Nonparametric maximum likelihood estimation of spatial patterns. *Ann. Statist.* **10** 1125–1136.

[20] SAGER, T. W. (1986). An application of isotonic regression to multivariate density estimation. In *Advances in order restricted statistical inference* (R. L. Dykstra, T. Robertson, and F. T. Wright, Eds.). Springer-Verlag.

[21] WEGMAN, E. J. (1970). Maximum likelihood estimation of a unimodal density. *Ann. Math. Statist.* **41** 457–471.

[22] DAHLHAUS, R. (1988). Empirical spectral processes and their applications to time series analysis. *Stoch. Process. Appl.* **30** 69–83.